LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-658098

Annual Report on characterization of genetic variability and virulence mechanisms of Venezuelan equine encephalitis viruses for DTRA

C. Jaing, J. Allen, N. Be, S. Gardner, K. McLoughlin, S. Weaver, N. Forrester, M. Guerbois

August 1, 2014

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-TR-658098

Annual Report on Characterization of genetic variability and virulence mechanisms of Venezuelan equine encephalitis viruses for DTRA

Contributors

Jonathan Allen

Nicholas Be

Shea Gardner

Kevin McLoughlin

Lawrence Livermore National Laboratory (LLNL), Livermore, CA

Scott Weaver

Naomi Forrester

Mathilde Guerbois

University of Texas, Medical Branch

Principal Investigator and Correspondent

Crystal Jaing

925-424-6574, jaing2@llnl.gov

Submission date: July 29, 2014

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Introduction

The overall objective of this proposed work is to characterize phylogenetic and phenotypic relationships of a diverse collection of Venezuelan equine encephalitis viruses (VEEV) to better understand epidemic amplification and virulence mechanisms. We will use a combination of high resolution genome-wide SNP microarray and deep DNA sequencing technologies against a panel of 200 VEEV isolates to discover genetic variations and understand VEEV evolution and phylogeny.

Venezuelan equine encephalitis virus (VEEV) is a mosquito-borne alphavirus capable of causing large outbreaks of encephalitis in humans and horses. Major epidemics dating to the early 20th century have affected hundreds of thousands of people and economically important equids. VEE complex viruses are endemic to South and Central America, Mexico, and Florida (Weaver et al. 2004). Although the fatality rate of VEEV is low in human infections (usually less than 1%), infection is typically highly debilitating and often results in permanent neurological sequelae (Quiroz et al. 2009). Moreover, because the disease primarily occurs in isolated rural areas and typical infections initially present with flu-like symptoms, many cases may go undiagnosed or are mistaken for other febrile diseases such as dengue (Vilcarromero et al. 2010). Unlike many naturally emerging arboviruses, VEEV is also a bioweapon threat due to its highly infectious aerosol-borne transmission history, its highly debilitating nature with very few inapparent infections, and its proven history as an effective weapon as developed during the cold war by the U.S. and former U.S.S.R. (Hawley and Eitzen 2001; Bronze et al. 2002) and its ability to efficiently express foreign genes that could render it even more virulent.

Enzootic VEE is also of particular concern due to its high burden of endemic human disease. Recent studies suggest that infected humans develop viremias sufficient in magnitude and duration to mediate transmission by the highly efficient urban vector, *Aedes aegypti* (Weaver and Reisen 2009). This scenario raises the possibility that VEE could, like dengue, become an urban disease throughout the Americas with even higher morbidity. For U.S. war fighters engaged in a conflict in Latin America, either direct exposure to the enzootic cycle in rural regions, as has been documented in Panama (Johnson et al. 1968; Quiroz et al. 2009), Colombia (Ferro et al. 2008), and Mexico, or infections in urban settings like Iquitos, Peru (Forshey et al.; Vilcarromero et al.; Watts et al. 1997; Watts et al. 1998; Aguilar et al. 2004; Vilcarromero et al. 2009; Vilcarromero et al. 2010) could inflict direct casualties and severely compromise their ability to fight.

There are three major challenges that we believe can be solved using new approaches: 1) Rapidly estimating the origin of a newly discovered VEEV strain; 2) estimating its equine and/or human amplification potential; and 3) Determining the human virulence phenotype of a newly discovered VEEV strain. Here, we propose to exploit high-throughput technologies that enable in-depth genomic sampling and characterization of a large strain panel to fill these crucial needs.

We used a combination of high resolution genome-wide SNP microarray and deep DNA sequencing technologies against a diverse panel of 200 VEEV isolates to discover genetic variations and understand VEEV evolution and phylogeny. The 200 isolates include: Virulent vs. a-virulent and epidemic amplification-competent vs. incompetent isolates; lab passaged vs. non-passaged isolates; geographically diverse isolates from South America, Central America, Mexico, Florida and Texas; and isolates from diverse hosts and from human outbreaks. We compared the genetic variability and the effect of lab manipulation by serially passaging two VEEV strains in two different cell lines, and analyzing the sequence differences by SNP microarray and Illumina sequencing. We will also use computational algorithms to identify amino acid variations in the VEEV proteins that will distinguish epidemic amplification-competent and virulent strains. Finally, we will validate the predicted genetic

markers using experimental *in vitro* assays with primary equine and human cells, and via mosquito infection assays using reverse genetics approaches.

Methods

Whole genome SNP analysis and microarray probe design

SNP analysis

We applied the kSNP software to find single nucleotide polymorphisms (SNPs) in whole genome data (Gardner and Slezak 2010). This is an alignment free method based on k-mer (oligos of length k) analyses. A SNP locus is defined by the sequence context of length k surrounding the SNP (k-1)/2 bases either side of the SNP with a variant SNP allele at the central base. SNP analysis was performed with k=13. This representation of a SNP locus is based on surrounding sequence information rather than positional information in a genome. It differs from traditional alignment-based concepts of a SNP locus, and it allows us to consider draft genomes which are available only as contig fragments in which positional information relative to the complete genome is not known. kSNP is also useful for viruses in which there may be highly divergent and poorly alignable regions among a large group of sequences, and conserved regions only exist among small subgroups of sequences. There is no bias that otherwise results from the choice of a reference sequence or from considering only a subset of regions of the genome that can be easily or quickly aligned. kSNP scales to hundreds of bacterial or viral genomes, and can be used for finished and/or draft genomes available as unassembled contigs. The method is fast to compute, finding SNPs and building a SNP-based phylogeny in seconds to hours. SNP-based trees can be calculated using parsimony, maximum likelihood (ML), or neighbor joining (NJ) on a distance metric of the number of SNP allele differences between each target sequence. SNP alleles were mapped to the nodes of the tree. kSNP detected 7926 SNP loci from the VEE genomes. After including 4 EEEV genomes as an outgroup, the total number of SNPs was increased to 9486 loci.

Probe Design

Microarray probes were designed for every SNP. Probe design strategy maximized sensitivity and specificity based on extensive prior lab testing on a Roche NimbleGen microarray platform, where we demonstrated 99.52% SNP allele call rates and 99.86% accuracy (Gardner et al. 2013). After testing seven alternative probe design strategies, we determined that maximum sensitivity and SNP discrimination accuracy result if the SNP base is at the 13^{th} position from the 5' end of the probe (the end farthest from the array), probes are between 32 and 40 bases long, and length varies so as to equalize hybridization free energy (ΔG) to the extent possible within the allowable length range. Probes shorter than 32 bases have high false negative rates, and longer probes are inefficient at discriminating single base mismatches. We found that ΔG is a better predictor of hybridization than T_m. Probe candidates with hybridization free energy below ΔG=-43 kcal/mol were shortened until either their ΔG exceeded -43 kcal/mol or they reached the minimum 32 bases. Probes were designed around the SNP on both the plus and minus strands, for all four possible SNP alleles, and all surrounding sequence variants.

We design probes for both the plus and minus strands; these are not the reverse complements of one another because the SNP does not lie at the center of the probe. There are probes for all observed

variants on each strand, so at least four probes per SNP locus for biallelic SNPs. In addition, any sequence variation outside of the k-mer SNP context of conserved bases is captured in multiple alternative probes for that allele, so there may be more than 4 probes per SNP locus, although for a given hybridization, only the probe variant with the best signal is used for assessing the SNP allele at the 13th position. Finally, probes are trimmed from the 3' end to remove any N's or other degenerate bases, and omitted altogether if doing so results in a probe less than 32 bases. If a probe is a subsequence of any other, only the shorter of the two is kept. For the VEEV data, including the subset of probes to detect only observed alleles in the available genomes required 70% fewer probes than would be necessary to include probes for the unobserved variants as well, allowing us to fit probes for all the SNP loci on a single 12x135K Roche Nimblegen array format, including duplicates for 89% of the probes. The probes on the array as well as the full set representing unobserved allele variants are available as supplementary data if needed.

Comparing whole genome versus single gene trees

SNPs from the E1, E2, E3, and capsid genes were extracted for separate analysis by identifying those SNPs that occurred within the specified gene region (Table 1). Parsimony trees were built from the MSA, all SNPs, and SNPs in each gene. These were compared in terms of the number of splits shared between different trees, calculated using CompareTree.pl (http://meta.microbesonline.org/fasttree/treecmp.html), and visualized with tanglegrams generated by Dendroscope. Equivalent branch rotations which did not change the relationships within a tree are performed by an algorithm to minimize the number of crossing lines between trees (Venkatachalam et al. 2010).

Table 1. Gene regions from which SNPs were extracted

Gene	SNPs between positions	Number of SNPs
E1	10000-11327	1268
E2	8563-9843	1384
E3	8386-8574	262
Capsid	7562-8396	937
All SNPs	1-~11500	9846

Select, extract, and produce cDNA from VEEV isolates.

We identified a subset of 136 representative strains based on temporal and geographic range, outbreak association, and prior genotyping data generated at UTMB (Table 2). Cultures of Vero monkey kidney cells were infected at a low multiplicity of infection to produce the passaged isolates, and the culture fluid was harvested when cytopathic effects were observed. The virus was precipitated using polyethylene glycol and NaCl, then centrifuged for concentration. RNA was extracted using Trizol (Life Technologies) according to the manufacturer's protocol. cDNA production was carried out using a mixture of random hexamer and dT oligo priming. This method yielded the best cDNA coverage along the genome, providing adequate coverage for subsequent comprehensive deep sequencing and microarray analysis.

Table 2. VEEV strains analyzed by SNP microarray.

Subtype	Strain	Passage History	Year of Collection	Host	Location Collected

IAB	111-73	sm3	1973	hor	Peru
IAB	69Z1	sm2,BHK1	1969	hum	Guatemala
IAB	Beck_Wycoff	sm8, cec1	1938	hor	Aragua St., Venezuela
IAB	CoAn5384	sm2,cec1	1967	hor	Cali, Colombia
IAB	E541_73	sm1, cec2	1973	hum	Guajira, Zulia State, Venezuela
IAB	Piura	sm3	1942	mule	Piura, Peru
IAB	TRD	sm7	1943	don	Trinidad
IAB	V-263E	u	1943	don	Trinidad
IC	12.225	V2	1995	hum	Venezuela
IC	12.563	V1,BHK1	u	hum	Venezuela
IC	6803	V1	1999	hum	Falcon State, Venezuela
IC	9813	V1	1999	hum	Urdueta, Lara State, Venezuela
IC	25716	BHK1	u	u	Venezuela
IC	25717	BHK1	u	u	Venezuela
IC	125567	BHK1	1997	hum	Zulia State, Venezuela
IC	243938	V1,BHK1	1996	hor	Trujillo State, Venezuela
IC	255005	SM3	2000	hor	Barinas State, Venezuela
IC	255058	sm2,V1	2000	hor	Carabobo State, Venezuela
IC	369673	V1	1999	hum	Manaure, Guajira, Colombia
IC	369676	V1	1999	hum	Manaure, Guajira, Colombia
IC	369678	V1	1999	hum	Manaure, Guajira, Colombia
IC	369680	V1	1999	hum	Manaure, Guajira, Colombia
IC	12.399	u,BHK1	1995	hum	Venezuela
IC	6119_gi20800451	V1	1995	hum	Falcon State, Venezuela
IC	INH9813	u	1995	hum	Urdueta, Lara State, Venezuela
IC	PHO127	BHK1	1962	hum	Guajira, Venezuela
IC	PHO1275	sm1,V1?	1962	hum	Guajira, Venezuela
IC	PMCHo5_gi20800448	u	1964	hum	Monagas, Venezuela
IC	SH3_gi5442468	V1	1993	hum	Candelaria, Venezuela
IC	SH5	V1	1997	hum	Candelaria, Venezuela
IC	V178	sm1, V1	1961	hor	Cundinamarca, Colombia
IC	V198_gi18152933	cec2	1962	hum	
IC	V202	sm1,V1	1962	hum	Guajira, Colombia
IC	ZGH734	V1	1999	hum	Sinamaica, Venezuela
IC	ZGH868	V1	1999	hum	Sinamaica, Venezuela
ID	247168	V2	2010	hor	Panama
ID	247186	V2	2010	hor	Panama
ID	309752	cec1	1974	hum	Lozania, Colombia
ID	312714	V2	1978	rat	Pto. Boyaca, Colombia
ID	980019	V1	2002	ham	Bosque San Miguel, Colombia
ID	980027	V1	1998	ham	Bosque San Miguel, Colombia
ID	980267	V1	2002	ham	Puerto boyaca, Colombia

ID	980408	V1	1998	mos	Casanare, Colombia
ID	980517	V1	2003	mos	San Pedro de la Paz, Colombia
ID	00SMH264	none	2000	ham	Monte San Miguel, Colombia
ID	00SMH279	none	2000	ham	Monte San Miguel, Colombia
ID	00SMH290	none	2000	ham	Monte San Miguel, Colombia
ID	00SMM515-11C	none	2000	mos	Monte San Miguel, Colombia
ID	02_2720_98	C6/36-1	1998	hum	Iquitos, Peru
ID	204381	u	1973	mos	Delta Amacuro, Venezuela
ID	212857	SMB-1	2003	hum	Darién, Panama
ID	213391	SMB-1	2003	hum	B del Toro, Panama
ID	23647	V1, BHK1	1974	ham	Catatumbo, Venezuela
ID	242959	u	1966	u	Gamboa, Panama
ID	249443_Yumare	sm2,V6	1972	ham	Yumare, Venezuela
ID	251641	sm3,V3	1976	ham	Pto. Concha, Venezuela
ID	307537	V1	1971	mos	Pto. Boyaca, Colombia
ID	309506	V1	1973	ham	Pto. Boyaca, Colombia
ID	334250	V2	1977	mos	Pto. Boyaca, Colombia
ID	335733	none	1978	ham	Pto. Boyaca, Colombia
ID	3880_gi323706	u	1961	hum	Canito, Panama
ID	474590	V2	1997	hum	P. metro, Panama
ID	481460	V2	2000	u	Peste, Panama
ID	4840	BHK1,sm2,V1,cec1	1961	hum	PA
ID	484551	V2	2001	u	Darien, Panama
ID	485029	V2	2001	u	Darien, Panama
ID	622-41	none	2000	mos	Monte San Miguel, Colombia
ID	75D143	cec1	1975	mos	Iquitos, Peru
ID	76V2561_1	sm4	1975	mos	u
ID	8138	cec2	1962	hum	El Rincon, Panama
ID	903104	BHK1	1977	mos	Bayano, Panama
ID	92CO-59	none	1996	ham	Los Corales, Colombia
ID	97Co42	v1	1997	ham	Monte San Miguel, Colombia
ID	98-003	u	2002	ham	Los Corales, Colombia
ID	98-007	u	2002	ham	Los Corales, Colombia
ID	993MM304-1	none	1999	mos	Monte San Miguel, Colombia
ID	CoAn59145	BHK3	u	ham	Tibu, Colombia
ID	CoAn9004_1	sm3, V1	1969	ham	Tumaco, Columbia
ID	FPI3700	V1	u	hum	Peru
ID	FSE507	V1	2000	hum	Iquitos, Peru
ID	FSL2314	u	2006	u	Loreto, Peru
ID	FSL2649	u	2006	u	Loreto, Peru
ID	FVB0204	u	2006	u	Cochabamba, Peru
ID	FVB0258	u	2007	u	Cochabamba, Peru
ID	GML903843	V1, BHK1	1984	hum	Bayano, Panama

ID	IQT1724	V1	1995	hum	Loreto, Peru
ID	MAC10	V1	1998	ham	Padron Agric. Station, Miranda State, Venezuela
ID	P676_gi14549692	u	1963	mos	
ID	Pan34958	p2,sm1	1976	mos	Catatumbo, Zulis State, Venezuela
ID	R16880	sm4, V1	1976	ham	u
ID	V209A	sm2,V2	1960	mus	u
ID	ZPC10	V1	1997	ham	Venezuela
ID	ZPC727	none	1997	ham	Las Nubes, catatumbo, venezuela
ID	ZPC820	none	1997	ham	Las Nubes, catatumbo, Venezuela
IE	2177B	V3	1968	u	Nicaragua
IE	63A216	sm1	1963	mos	Veracruz, Mexico
IE	63Z1	sm1, V1	1963	hum	Veracruz, Mexico
IE	65U206	sm1	1965	ham	Sontecomapan, Mexico
IE	66U91	cec1	1966	ham	Sontecomapan, Mexico
IE	67U201	sm1	1967	ham	Belize
IE	67U208	V?	1967	ham	Honduras
IE	67U222	V1	1967	ham	Minititlan, Mexico
IE	67U225	sm1	1967	ham	Pto. Cortez, Honduras
IE	68U200	none	1968	ham	La Avellana, Santa Rosa Department, Guatemala
IE	68U201	u,sm1	1968	ham	La Avellana, Guatemala
IE	68U217	BHK1	1968	ham	Pto. Barrios, Guatemala
IE	69U315	sm1	1969	ham	Sontecomapan, Mexico
IE	70U1134	u	1970	ham	Iquitos, Peru
IE	70U74	u	1970	ham	Pto. Barrios, Guatemala
IE	71U382	V1	1971	ham	La Avellana, Guatemala
IE	71U384	sm1, V1	1971	ham	Santa Rosa Department, Guatemala
IE	72U23	V1	1972	ham	La Avellana, Guatemala
IE	73U151	u	1973	ham	La Avellana, Guatemala
IE	78U202	u	1978	ham	La Avellana, Guatemala
IE	79U13	BHK 1	1979	ham	Izabal Department, Guatemala
IE	80U76_gi17865005	u	1980	hor	La Avellana, Guatemala
IE	BT2607	u	1961	mos	Almirante, Panama
IE	MenaII_gi4262302	u	1962	hum	Zulia State, Venezuela
IE	MX01-32	none	2001	ham	Chiapas State, Mexico
IE	MX03H1	none	2003	ham	Las Coaches, Pijijiapan, Chiapas State, Mexico
IE	MX08H50	V1	2008	ham	E. Coachapa, Minatitlan, Veracruz State, Mexico
IE	MX08H53	V1	2008	ham	Tacoteno, Minatitlan, Veracruz State, Mexico
IE	MX09M64	V1	2008	mos	Tacoteno, Minatitlan, Veracruz State, Mexico

IE	MX10_94M5	none	2010	mos	Minatitlan, Veracruz State, Mexico
IE	MX10H91_00011	none	2010	ham	Minatitlan, Veracruz State, Mexico
III	FSL0190	V2	2000	hum	San Juan, Iquitos, Peru
III	PC254	u	1997	rat	Iquitos, Peru
III	PE407660	u	1998	mos	Iquitos, Peru
IIIB	MucamboBeAn8_gi4262305	u	1954	u	Brazil
IV	PixunaBeAr35645_gi4262314	sm4, V1	1961	mos	Brazil
V	CabassouCaAr508_gi4262323	sm10	1968	mos	French Guiana (Cabassou)
IC	243937_gi5442464	V1, BHK1	1992	hor	Trujillo State, Venezuela
IC	254934_gi62824833	BHK1	2000	hor	Barinas State, Venezuela
IC	255010_gi62836644	sm2,V1	2000	hor	Barinas State, Venezuela
IE	MX03H2	none	2003	ham	Las Coaches, Pijijiapan, Chiapas State, Mexico
IE	MX09Eq03	V1	2009	hor	Tacoteno, Minatitlan, Veracruz State, Mexico

Extract whole RNA from infected animal tissue.

Mouse brains infected with VEEV vaccine strain TC83 were subjected to whole RNA extraction. Brains infected with Chikungunya virus vaccine strain 181/clone 25 were examined in parallel for control purposes. Three biological replicates were examined for each viral infection. Brain tissue was homogenized in Trizol (Life Technologies). RNA was extracted and purified using the Direct-zol RNA MiniPrep kit (Zymo Research) according to the manufacturer's instructions. Whole cDNA was synthesized as described above.

Fabricate 12-plex 135K NimbleGen arrays and hybridize VEEV cDNA to the arrays.

All cDNA samples were fluorescently labeled and hybridized to VEE SNP array as described (Jaing et al. 2008). Briefly, fluorescent labeling of samples was performed using the Nimblegen One-Color DNA Labeling Kit (Roche). One µg VEEV cDNA was added to Cy-3 labeled random primers, followed by isothermal amplification at 37°C using klenow polymerase. Labeled DNA was purified via isopropanol precipitation and resuspended in water for microarray hybridization. DNA samples were prepared for hybridization using the Nimblegen Hybridization Kit LS. Three µg labeled DNA were hybridized to each array, followed by incubation for 40-45 hours at 42°C. Arrays were washed using the Nimblegen Wash Buffer Kit. The fluorescent signal on the array was scanned using a 2 µm Roche fluorescent scanner MS200. The array raw data was generated using the NimbleScan software available from Roche NimbleGen.

Examine phylogenetic relationships between and evolution of VEEV strains based on SNP microarray data.

We used our previously developed analysis software to call alleles at each locus for each sample analyzed on SNP microarrays. The software fits a linear model of strand and allele effects to the log intensity data from all probes for the locus, and calls the allele as the one with the largest coefficient in the fitted model. Separating the strand and allele effects is necessary in order to compensate for the differing hybridization efficiencies often seen between forward and reverse strand probes.

Because our definition of a SNP locus requires conservation of the 6 bases on either side of the polymorphic base, array probes for one locus may hybridize to genomes in which a similar locus context is present. That is, loci that are considered to be different in the sequence analysis, but have 13-mer contexts that are identical except at one or two positions, may be difficult to distinguish by microarray probes. Therefore, our current array analysis software does not try to determine whether a locus is present or absent; i.e. an allele call is made for every locus.

For isolates analyzed on the array that had genome sequences available, we computed the concordance rate between the allele calls from the array and the genome sequence, as the fraction of loci present in the genome for which the array calls agreed. We also computed the numbers of allele differences between each array sample and each genome, and determined whether the closest genome was in fact the genome sequence for that strain.

We used the combined genotype data from SNP microarrays and genome sequences to create maximum parsimony phylogenetic trees, using Parsimonator (https://github.com/stamatak/Parsimonator-1.0.2.). We chose the best (most parsimonious) of 100 trees generated using different random number seeds.

Phenotype/genotype associations

We identified variable positions in the MSA that were non-randomly associated with a given subtype or host according to chi-squared tests using PPFS (https://sourceforge.net/projects/ppfs) (Hall 2014). Using 100 bootstraps per phenotype, we identified positions most reliably associated with each binary phenotype across bootstrap replicates. For each position/phenotype association, we counted the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), and used them to compute the accuracy = (TP + TN) / (TP + FP + TN + FN), positive predictive value (PPV) = TP/(TP + FP), negative predictive value (NPV) = TN/(TN+FN), true positive rate TPR=TP/(TP+FN), and true negative rate TNR=TN/ (TN+FP). We also reported the number of positions with chi-square $p < 7.98 \times 10^{-8}$, based on the Bonferroni correction with a single test p value of 0.001 divided by 12524, the number of non-consensus positions in the alignment.

Perform whole genome sequencing of a subset of VEEV RNAs to confirm microarray data.

We selected a subset of 10 isolates for whole genome sequencing using the 10 barcode multiplex Illumina sequencing technology to confirm microarray findings. We sequenced at a minimal depth of 100x to identify SNPs and other variations among different VEEV samples. The quality values of individual base calls and the alignment scores of individual reads were evaluated using a collection of open source software tools designed to recover high confidence SNP calls while minimizing the potential for false negatives.

We applied a modified version of our existing bioinformatics software package designed for characterizing the mutant spectra of a viral population using Illumina sequencing data (Chen-Harris et al. 2013). Reads are mapped to a near neighbor genome, which is selected from the VEE whole genome database when the isolate could be matched with the newly sequenced sample. When there was no clear near neighbor, the reads were mapped to all VEE genomes in the database and the genome with the highest number of mapped reads was selected. A new “consensus” genome is reconstructed from the mapped reads and the reads were then re-mapped to the new consensus

genome. This process is repeated until the newly assembled genome converges on a single consensus sequence. The software applied a statistical framework to report predicted rare variants that can be confidently distinguished from mutations introduced by sequencing error by counting the number of reads that map to the consensus genome (or dominant genotype) and contain a rare variant and occur with sufficient frequency to be unlikely explained by sequencing error. There were two modifications to the published protocol since we did not have access to overlapping read-pairs to do exact error correction and sequencer error modeling. Therefore, we relied on earlier work to apply previously quantified error profiles.

Characterize genetic variation in two passaged VEEV strains vs. their natural unpassaged strains.

We selected two phylogenetically distinct unpassaged isolates, one from a mosquito pool (MX10-94M5) and one from a sentinel hamster (00SMH279), performed ten serial passages of each in Vero and C6/36 mosquito cells then performed Illumina deep sequencing of the four passaged virus populations with the sequencing of the two unpassaged samples currently in progress. Three of the four sequenced passage samples were successfully completed with the fourth sample (MX10-94M5 passage 10) currently being re-sequenced due to a low sequence data yield.

Table 3 shows the high coverage returned for each of three sequencing runs, which indicate the potential to detect rare variants down to 0.02% with 853,589x coverage. The number of rare variants present is reported to number from 2,823 (OSHMH279 C636) to 3,354 (OOSMH279 Vero), however, given the potential variation in sequencing error in the current experiments relative to previous estimates, this count could include some false positive calls. Table 3 also shows a more conservative estimate by counting rare variants that occur with > 0.1% frequency in the observed reads.

Infection of monocytes and analysis of VEEV replication patterns.

To examine differential patterns of viral replication between diverse VEEV strains, we proposed to perform monocyte infection assays. Isolation of human monocytes from fresh blood is a delicate and lengthy process, which requires a large amount of blood in order to obtain sufficient monocytes to complete infections with 10 VEEV strains. In order to simplify this process, we elected to use a monocyte cell line to complete this part of the task. THP-1 cells (ATCC TIB-202TM) were obtained and cultured according to ATCC guidelines. We performed the initial pilot experiment with 5 strains: IAB TrD, IC 3908, IC SH3, ID ZPC738, and IE 68U201.

THP-1 human macrophage cells were infected in triplicate at a multiplicity of infection of 10 for one hour in suspension, then transferred infected cells to a 12 well plate for incubation. 100 µl aliquots of supernatant were harvested at 0, 12, 24 and 48 hours post-infection. Supernatants were then analyzed by plaque assay to determine titers.

Table 1. Sequencing Coverage of 14 sequenced samples. The three samples with low sequencing coverage are being repeated. We'll submit an updated table once it is complete.

Strain designation	Consensus sequence length (>20x)	Rare variants (>0.1%)	Rare variants (all)	Coverage (x1000)	Name of reference strain
FSL0190	10572	63	2009	104	71D-1252
GML908408	8174	-	239	0.065 (Low)	Mesocricetus auratus/COL/97CO-42/1997/ID
MAC10	11385	223	2059	97	Mosquito/PER/75D143/1975/ID
MUCAMBO	8810	666	2208	50	Mucambo BeAn 8
PE40766	10486	67	1487	74	71D-1252
PIXUNA	5843	-	744	0.14 (Low)	Pixuna BeAr 35645
249443	11365	154	1393	92	Equus caballus/PER/Hoja Redonda/1971/IAB
58_73-IAB	11403	415	2478	361	Equus ferus caballus/PER/111/73/1973/IAB
CABOSSOU	11385	313	2350	82	Cabassou CaAr 508
CLH2293	10532	77	2784	251	71D-1252
00SMH279_C636_P10	11214	137	2823	187	Mesocricetus/auratus/COL/00SMH279/2000/ID
00SMH279_VEROS_P10	11215	110	3354	302	Mesocricetus/auratus/COL/00SMH279/2000/ID
MX10-94M5_C636_P10	2902	-	66	0.015 (Low)	Mosquito pool/MEX/MX10-94M6/2010/IE
MX10-94M5_VEROS_P10	11446	70	3251	295	Mosquito pool/MEX/MX10-94M6/2010/IE

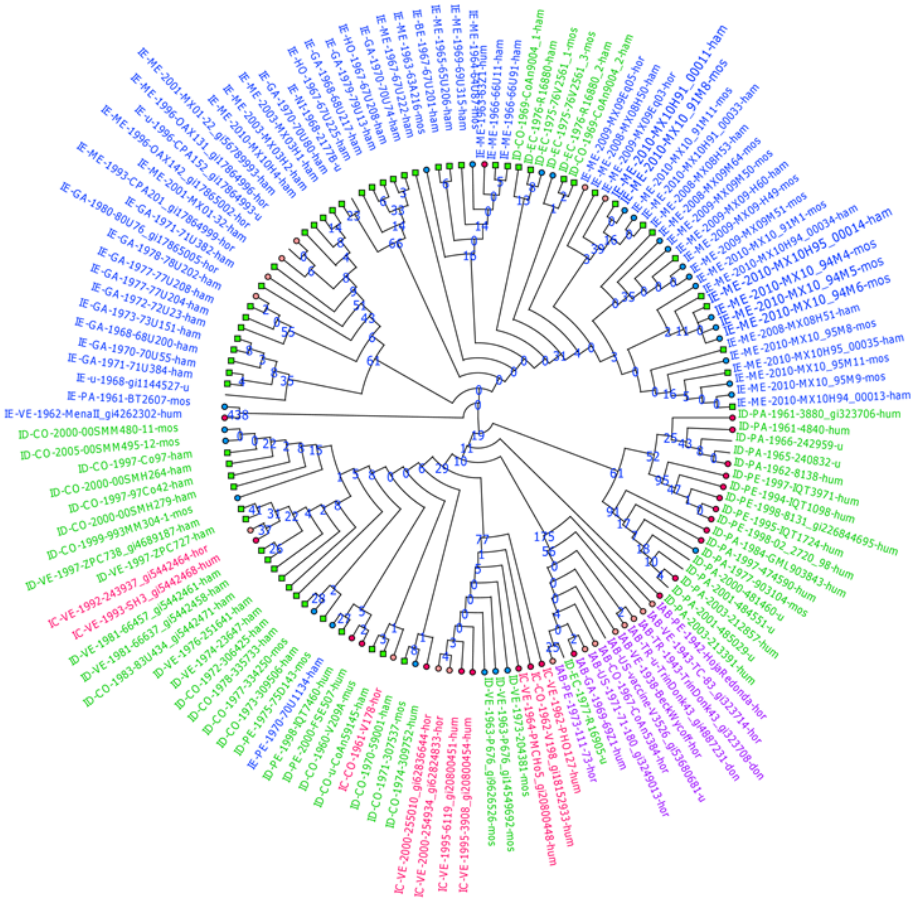
Results

Whole genome SNP analysis and phylogenetic tree construction

When we compared SNP-based trees built by different methods to a tree constructed from a whole genome multiple sequence alignment (MSA), the SNP tree built with parsimony (Figure 1) was more similar to the MSA-based tree than those built with NJ or ML. Out of all splits in the alignment-based tree, 77% were represented by splits in the parsimony tree, compared to only 68% in the ML tree. Moreover, the parsimony tree had fewer homoplastic SNPs than the ML tree (1679 versus 2153, respectively, from the dataset using the EEEV outgroup genomes). Homoplastic SNP loci are those in which the pattern of shared alleles does not conform to any of the branches of this tree, as a result of processes such as convergent evolution, homologous recombination, multiple mutations at the same site, or sequencing errors. Maximum parsimony has been shown to outperform ML in phylogenies that

display heterotachy, a phenomenon in which the rates at which different sites evolve change over time (Kolaczkowski and Thornton 2004). In this case, non-parametric estimation of trees by parsimony is more accurate than parametric methods such as ML. In VEEV, evolutionary rates may vary more over time in genes playing a larger role in transmission and immune evasion, such as those encoding the envelope and capsid, than in structural or polymerase genes. For this single stranded RNA virus, rapidly evolving genes are about a third of the genome, so that heterotachy has a strong effect on genome evolution. This could explain why the parsimony SNP tree seems to more closely resemble the MSA based tree and has fewer homoplasies. By contrast, when we build phylogenetic trees for bacteria with thousands of genes, rapidly evolving regions are a small fraction of the genome. In this case heterotachy has a smaller influence on genome evolution, and we usually find that ML SNP trees produce more accurate results than parsimony.

Figure 1. SNP phylogeny by parsimony. Strains labeled by subtype-country-year collected-strain-host. Country: GA=Guatemala, PE=Peru, NI=Nicaragua, VE=Venezuela, CO=Colombia, TR=Trinidad, PA=Panama, US=USA, EC=Ecuador, ME=Mexico, BE=Belize, HO=Honduras, BR=Brazil, AR=Argentina, FG=French Guiana. Host: hor=horse, don=donkey, hum=human, mos=mosquito, ham=hamster, mus=mouse. u=unknown. Subtype: blue=IE, green=ID, red=IC, and purple=IAB). Collection host: symbols at branch tips (red circles=human, orange circles= horses, blue circles=mosquitos, and green squares=hamsters). Number of shared alleles shown in blue at nodes.



Almost all VEEV strains could be uniquely identified by their genotypes across these SNP loci. Numbers at the nodes of the tree in Figure 1 indicate the number of loci at which the allele is uniquely shared by all and only the strains down that branch. Only two sets of genomes were unresolved (i.e., had identical genotypes across all 7,926 SNPs, Figure 1 strains in italic, not bold, type). One consisted of two genomes collected on successive days from Minatitlan, Mexico on August 26-27, 2010: one from a mosquito pool and the other from a sentinel hamster. The other comprised four genomes, also collected from Minatitlan in 2010; the first three collected from mosquito pools on August 26-27, and the fourth from a hamster on August 28. These results indicate that sentinel hamsters do become infected with the variants circulating in insect vectors in the area at the time. These isolates were members of a larger group of closely related genomes collected in Minatitlan, Mexico between July 2008 and late August 2010 in hamsters and mosquitos, as well as two from horses.

Tree organization with respect to serotype, collection date, and host

The phylogeny generated by whole-genome SNP analysis (Figure 1) has overall structure similar to a previously published phylogeny, which was based on alignment of the E2 gene and parts of the E3 and 6K genes from a smaller set of strains (Anishchenko et al. 2006a). However, the larger set of complete genomes used in our analysis makes it possible to resolve both high level clades and fine scale SNP differences among closely related strains. Several patterns emerged.

First, we extended previous results (Weaver and Barrett 2004) showing that strains with high overall similarity across the whole genome may exhibit different antigen serotypes. For example, the epizootic type IAB strains and associated vaccine strains (Figure 1) collected from multiple countries from 1938-1973 form a distinct clade of highly similar isolates; however, this clade also includes a subtype ID isolate (R16905) collected in 1977. Likewise, all of the epizootic/epidemic IC strains have high similarity to groups of enzootic type ID isolates. In turn, subtype ID isolates can be found in both of the two major branches of the phylogeny; one set that clusters with the type IAB and IC strains previously mentioned, and another group that appears to have emerged from the enzootic subtype IE strains that make up most of the upper branch of the tree. We also found one case of a type IE strain (70U1134) that groups with a set of ID isolates in the lower branch of the tree.

Second, when we examined the collection dates of samples found in each clade, we found that many clades were remarkably persistent. While a few clades were relatively transient, with collection dates all within a few years (e.g. the IE isolates sampled in Mexico from 2008-2010), most clades persisted over one or more decades. For example, the subtype IAB epizootic strains (and associated type ID outlier) showed little genetic variation, even though they were collected over nearly 40 years (1938-1977) across a wide geographic area, from USA through Guatemala and Trinidad down to Venezuela and Peru. Likewise, the subtype IC and ID isolates comprising the lower part of the tree in Figure 1, collected between 1961 and 2005, have very few differences across our panel of SNP loci.

Third, we found that phylogenetic groupings were not in general associated with particular hosts; the broad associations that do appear are likely artifacts of the different sampling strategies used for enzootic (subtype ID and IE) strains, which account for all samples from mosquitos and sentinel hamsters, and for epizootic (subtype IAB and IC) strains, which comprise most samples from equids and humans. Within the major branches of the phylogeny, we find many cases of human- or equid-infecting isolates that are closely related to strains collected from hamsters or mosquitos. For example, the only human-infecting genomes from enzootic subtype IE strains (IE-VE-1962-MenaII_gi4262302-

hum and IE-ME-1963-63Z1-hum) are each nearly identical to isolates collected from mosquitos or hamsters, though unrelated to one another. The serotype ID infections from Panama and Peru collected from 1961-2003 is the largest cluster of mostly human isolates (lower right quadrant of the tree), uniquely sharing 61 SNPs, 11 of which are nonsynonymous, landing on both the structural (e.g. E2) and nonstructural polyprotein (e.g. nsp3). There was only one fatal human infection, 3880 (gi|323706) from Panama in April 1961.

Phenotype prediction

Because the host and subtype do not correspond exactly to the phylogeny, this gives us the opportunity to search for positions associated with these important phenotypes and which are not simply a product of ancestry. We applied the PPFS package to identify variations that are associated with particular hosts or subtypes. Our results indicate that these phenotypes are complex polygenic traits affected by multiple alleles on multiple genes.

Phenotype prediction of subtypes displayed accuracies of 90% for ID up to 99% for IAB and IE (Table 4). The SNP at position 213 on the E2 protein shown previously (Anishchenko et al. 2006b) to mediate the shift from ID to IAB or IC (T213 -> K or R) was also identified here to associate non-randomly with phenotype, although the association was not as clear as previous studies had found; five ID strains had the K213 allele that Anishchenko et al. had concluded would convert the phenotype to IC (ID-EC-1977-R16905-u, ID-PA-1962-8138-hum, ID-VE-1963-P676_gi14549692-mos, ID-VE-1963-P676_gi9626526-mos, ID-VE-1973-204381-mos). IC and ID had lower accuracy, PPV, etc. than IE and IAB, as expected from the pattern of closely related mixed types in the tree (Figure 1). It was not possible to accurately distinguish many of the IC, ID, and the lone IE in the lower half of the tree in Figure 1 using the predictive models built by PPFS.

Table 4. Accuracy, positive and negative predictive value, true positive and negative rates, the number of variable positions used in the PPFS prediction model, and the number of variable alignment positions non-randomly associated with each phenotype.

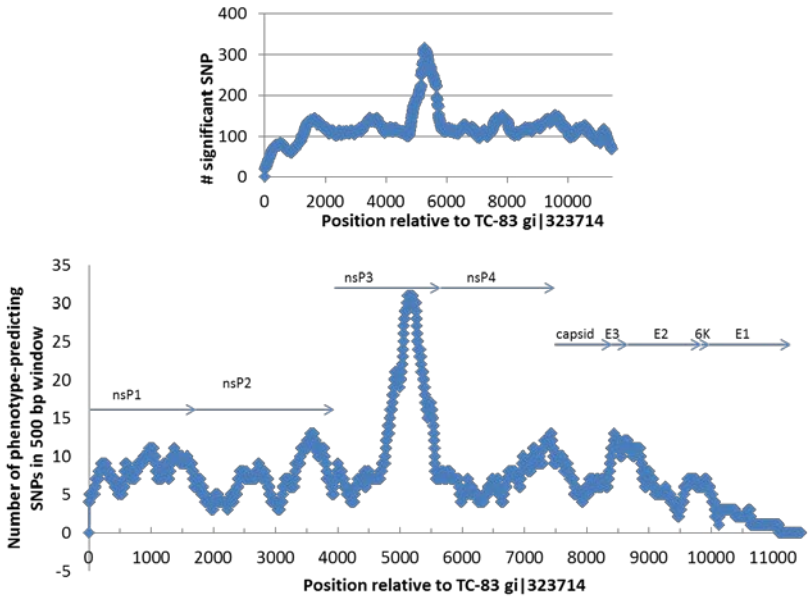
	Accuracy	PPV	NPV	TPR	TNR	# Align-ment positions used in prediction	# Align-ment positions signficant *
IAB	0.99	0.91	1.00	1.00	0.99	32	166
IC	0.96	0.70	0.98	0.70	0.98	54	42
ID	0.90	0.82	0.96	0.94	0.88	18	2305
IE	0.99	0.98	0.99	0.98	0.99	19	2584
hamster	0.75	0.65	0.93	0.95	0.59	17	0
horse	0.92	0.89	0.92	0.44	0.99	37	37
human	0.90	0.92	0.90	0.50	0.99	6	123
mosquito	0.82	0.67	0.85	0.45	0.93	5	0
horse or human	0.87	0.91	0.85	0.76	0.95	30	46

Accuracy of host prediction was lower than for subtype prediction, ranging from 75% for hamster up to 92% for horse. The TPR, i.e. the number predicted to be positive that actually are, for mosquito, horse, and human hosts was low, and the TNR for hamster hosts was low. Considering human and horse hosts combined as a single large mammal phenotype and counting mosquito hosts as unknown had only a minor affect, slightly improving TPR but not overall accuracy. Close inspection of the mutations identified as significant showed that they followed the phylogeny, and no mutations that universally associated with host or serotype across multiple different phylogenetic branches could be identified. This was the reason all our phenotype/genotype association models required up to several dozen positions, and still made false positive and negative calls. The 6 positions identified for predicting association with human hosts are all nonsynonymous, with 3 on the nsP3 protein, and one each on nsP4, capsid, and E2. While significant, these are not sufficient predictors of human outbreak potential, as shown by the low TPR in the predictive model.

Because of the non-random association of host and subtype (all hamster and mosquito samples were from subtypes ID and IE), we also stratified by type prior to searching for SNPs associated with host. No SNPs perfectly discriminated whether the host was human or horse in the IC or the combination of IC and IAB subtypes, suggesting that IC and IAB strains collected from horses do not consistently differ from those isolated in human hosts. Nor was it possible to identify SNPs in the IE subtype strains to perfectly distinguish those isolated from horse, hamster, or mosquito.

The nsP3 gene appears to be a hotspot for subtype/host-associated mutations (Figure 2). The main plot shows the density of variable positions selected for inclusion in the PPFS prediction model in a 500 bp sliding window, while the inset shows the number of significant positions including many that were not selected for the predictive model (because they provided equivalent or redundant information with those mutations already selected).

Figure 2. Density of variable positions in a sliding 500 bp window selected by PPFS for genotype/phenotype prediction, combined for all host and subtype models, plotted relative to the TC-83 genome, with horizontal lines showing the approximate positions of the mature peptides. The smaller inset plot shows the density of all significant positions with chi-square $p<7.98x10^{-8}$.



We also ran PPFS on the SNPs predicted by kSNP rather than the variable positions from an alignment, and found that PPFS predictions from variable positions in the MSA were slightly more accurate and generated models relying on fewer positions than predictions based on kSNP SNPs. VEE is highly variable, and mutations in the context around a SNP caused some homologous positions across strains to be considered as different loci by kSNP, since kSNP defines loci by conserved flanking sequence and does not consider indels. This resulted in large numbers of missing loci for each strain.

Comparing whole genome alignment to SNP analysis, and single gene SNPs to complete genome SNPs

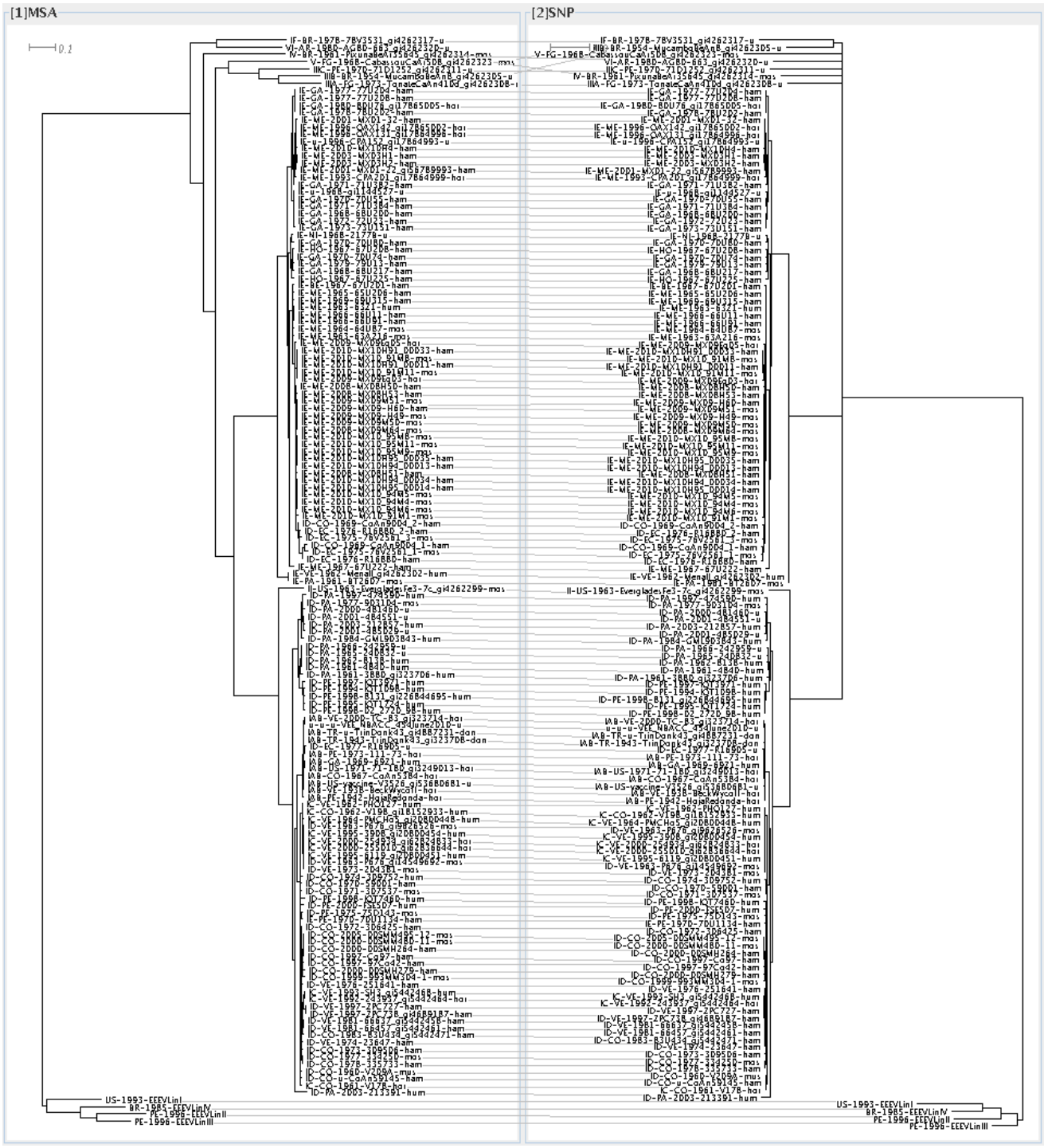
Comparing the tree from all SNPs versus the trees generated from using only the SNPs in a single gene shows that only about half the splits are present in the gene trees from any of the envelope genomes compared to the tree based on all SNPs in the genome (Table 5). This is not surprising, since only 13% of the SNPs fall on the E1 gene, so lower resolution and accuracy is expected. The capsid gene SNPs are somewhat worse, with only 37% of the splits observed, despite the fact that there are over 3.5 times more SNPs in the capsid gene than the E3 gene, which has the smallest number of SNPs for building the tree. The E1 gene results in a better representation of the tree than E2 or E3, as it captures almost 10% more of the splits identified from all the SNPs.

Table 5. Comparison of trees from MSA versus all SNPs, and trees from SNPs located in a single gene versus all SNPs.

Tree comparison	**Splits Found in 2nd tree**	**Total Splits in 1st tree**	**Fraction splits in 1st tree found in 2nd tree**
MSA vs all SNPs	112	146	0.77
All SNPs vs E1	84	146	0.58
All SNPs vs E2	72	146	0.49
All SNPs vs E3	68	146	0.47
All SNPs vs capsid	54	146	0.37

Figure 3 shows a tanglegram with the tree from the MSA on the left and all SNPs on the right, with lines connecting the same taxa between trees. Differences between these trees are minor and within a reasonable expectation of uncertainty in the trees, mostly around the poorly resolved Mucambo, CabassouCaAr, PixunaBeAn, etc. These were the strains collected from mosquito pools from 1954-1980 from geographically dispersed French Guiana, Brazil, Argentina, and Peru, and are now considered different species in the VEEV antigenic complex (Weaver and Barrett 2004). Each of these genomes has about 500 genome specific SNP alleles and they are a diverse set of V, IV, VI, IIIA, IIIB, IIIC, and IF species with only one representative each, each branching off the tree basal to the branches leading to the more heavily sequenced subtypes from Mexico, Peru, and Venezuela. In summary, the similarity between the whole genome SNP and MSA trees supports our SNP genotyping approach as a rapid, cost effective method to phylogenetically characterize unsequenced samples using SNP arrays.

Figure 3. Tanglegram connecting the corresponding taxa which illustrates the high similarity between the MSA tree (left) and the SNP tree (right). Tanglegrams were created with Dendroscope (Scornavacca et al. 2011).



Figures 4 and 5 show the tanglegrams with the tree from all SNPs on the left and the tree from the SNPs in the E1 (Figure 4) or capsid (Figure 5) gene on the right. The EEE genomes are not clustered as a monophyletic group in any of the SNP gene trees. Further, the capsid gene SNP tree has lower

accuracy than the envelope gene trees, illustrated by the many crossing lines of the tanglegram in Figure S3. The differences between the gene SNPs trees and the whole genome SNPs tree illustrate the difficulty of phylogenetic analyses based on a small region rather than the full length of the genome, and suggest that SNP phylogenies created from a single gene may risk low resolution and accuracy.

Figure 4. Tanglegram illustrating differences between the SNP tree based on all SNPs (left) and the tree based only on the SNPs in the E1 gene (right).

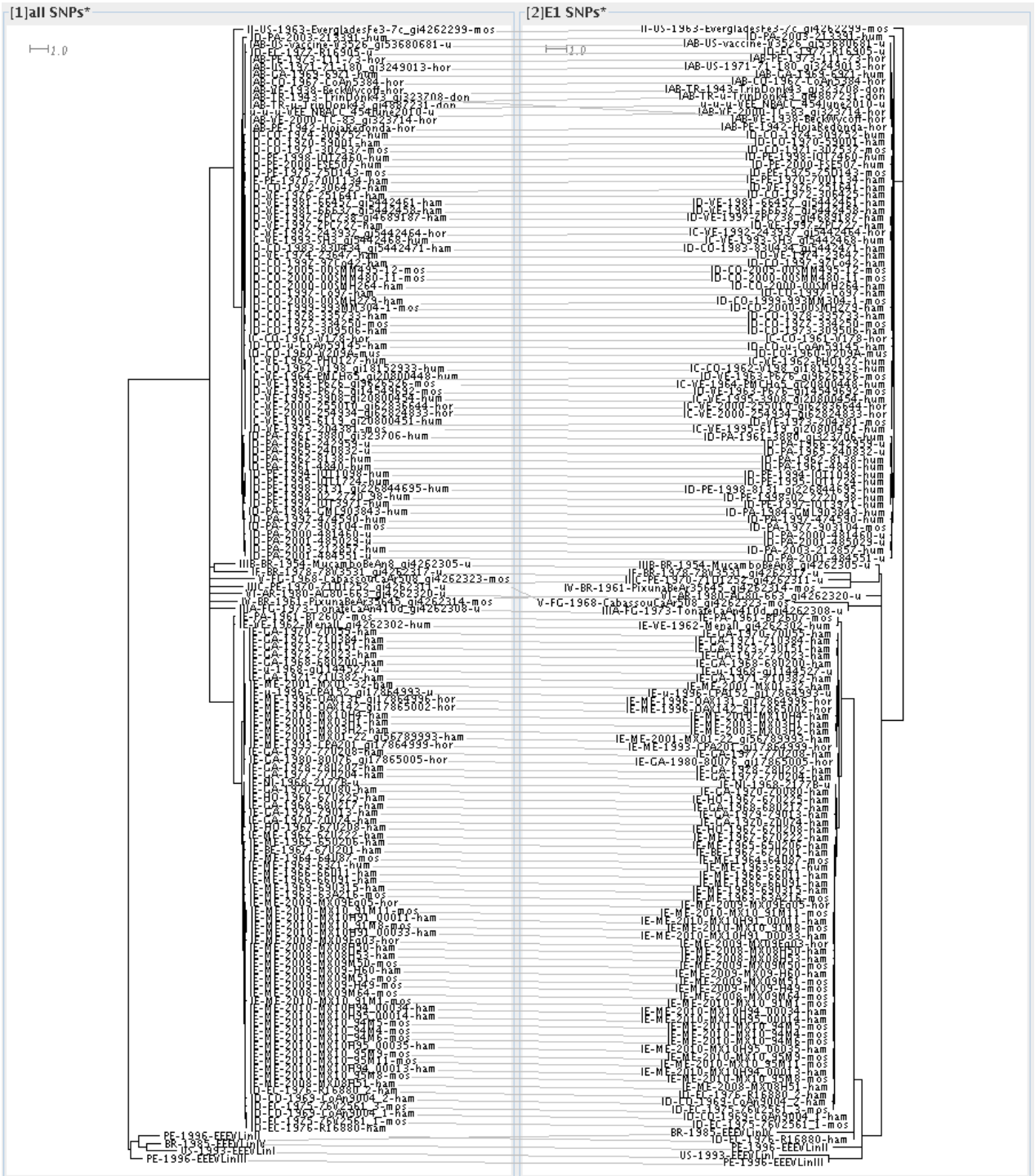
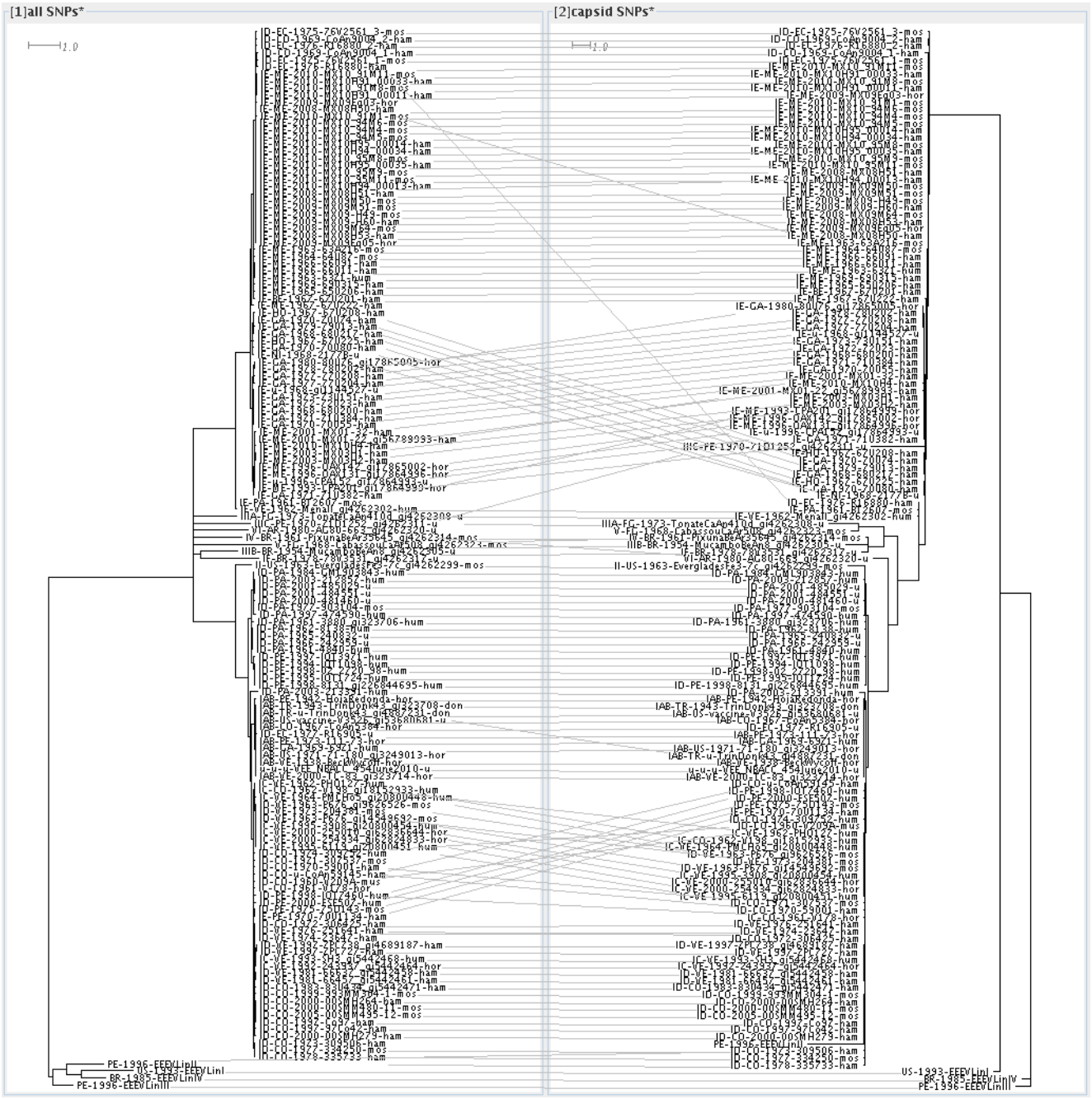


Figure 5. Tanglegram illustrating differences between the SNP tree based on all SNPs (left) and the tree based only on SNPs in the capsid gene (right).



Microarray analysis of VEEV cDNA samples

We hybridized cDNAs from 136 isolates to SNP arrays. Genome sequence data was available for 82 of the samples. The overall concordance rates were calculated between the allele calls made by SNP microarray versus those called by whole genome sequence data. The overall concordance rate was 97.47%. Hybridizations of replicate cDNA samples extracted from the same isolate showed close agreement between replicates. One source of error was that the array analysis currently is not able to call a locus as missing, even if that locus is not present in the genome sequence, causing discordance between the genome and array. The array correctly

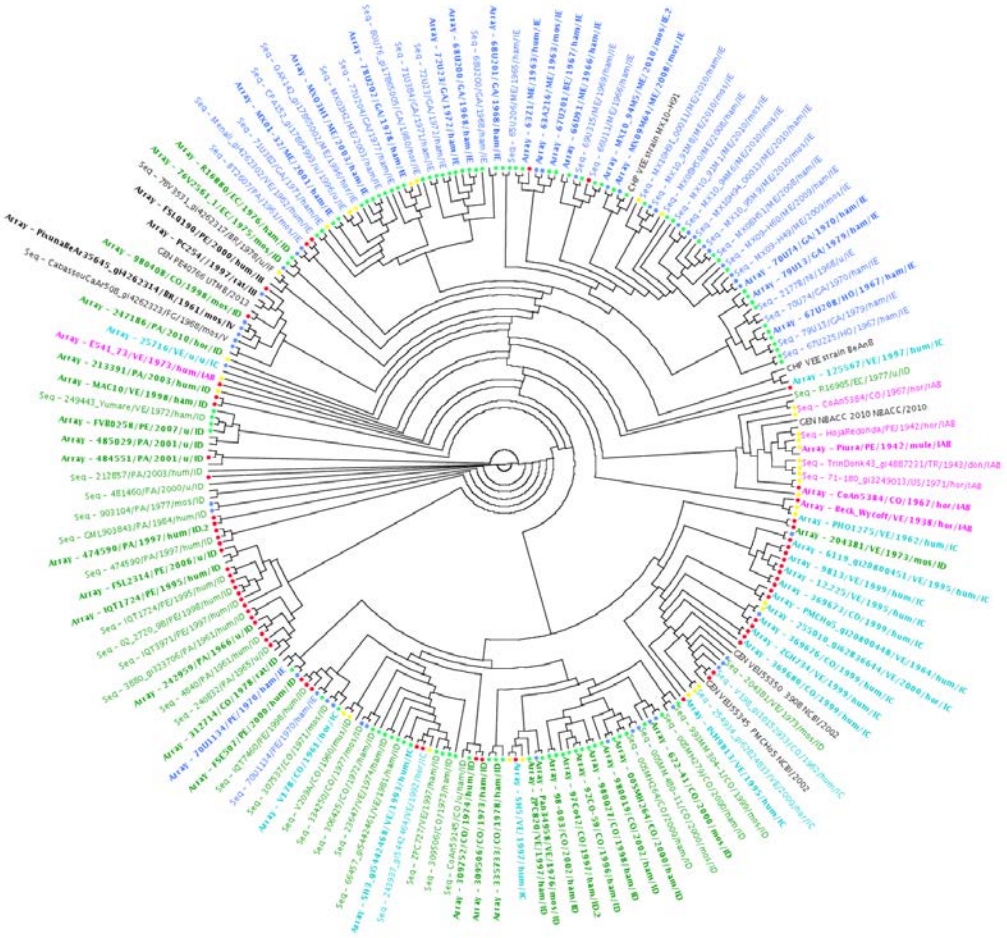
classified 70 out of 75 cDNA samples. The 5 misclassified cases (Table 6) were to highly similar sequences collected in the same location, some of which are arguably the same strain (e.g. MX03H1 and MX03H2), and the concordance rates were almost identical between the correct and incorrect strain.

Table 6: Strains where the closest genome called by the array was incorrect. Each pair of sequences lists the genome sequence followed by the array strain call.

Strain	host	Subtype	Country	Year
255010gi62836644	hor	IC	VE	2000
254934gi62824833	hor	IC	VE	2000
PE407660	mos	III	PE	1998
FSL0190	hum	III	PE	2000
MX03H2	ham	IE	ME	2003
MX03H1	ham	IE	ME	2003
MX09M51	mos	IE	ME	2009
MX09M64	mos	IE	ME	2008
MX1091M8	mos	IE	ME	2010
MX10H9100011	ham	IE	ME	2010

A parsimony based phylogenetic tree was generated using both SNP microarray and whole genome sequencing data (Figure 6). The SNP array data are shown in bold and labeled as "Array", and the whole genome sequence data is shown in plain text and labeled as "Seq." Serotypes are color coded (ID in green, IC in aqua, IAB in plum, IE in blue). Host from which the sample was collected are color coded by the circle at the tips (human in red, horse and donkey in yellow, mosquito in blue, hamster in light green). Array and sequence data for the same strain appear in close proximity on the tree, especially for the ID and IE subtypes. Some of the IC array isolates cluster together and are not nearest neighbors with their respective genome sequence, although both the array samples and genomes fall under a very closely related branch. The IC strains that diverge from this large IC cluster, and instead appear down branches of mostly ID isolates, do place the array result adjacent to the genome data from the same isolate. Thus, differences of a just a few SNPs among highly similar isolates are better resolved by genome sequencing, while array data provides sufficient accuracy in phylogenetic classification to correctly cluster isolates by clade and to identify the closest neighbors that have been sequenced or hybridized to the array. Array data is suitable to classify isolates by subtype, host, country, and year to the extent that these correlate with phylogeny.

Figure 6. Phylogenetic tree of SNP array data and whole genome sequence data of VEE, using parsimony.



Microarray analysis of infected tissue samples

To determine if strain genotyping could be performed directly from infected tissue, we extracted whole RNA from infected mouse brains, produced cDNA, and processed these samples on the SNP array. The tested strain (TC-83) was successfully identified in each of three samples. Successful typing directly from samples that would potentially be collected in the field could allow for strain identification without the time consuming and laborious steps inherent in isolation and culture.

Whole Genome sequencing of a subset of VEEV strains

We compared genetic variation among the dominant genotype of a subset of VEEV strains (i.e. the consensus assembled genomes), comparing the unpassaged and the subsequent passaged samples. Table 7 summarizes the 11 genome positions where there was a mutation observed between the unpassaged and passaged 00SMH279 samples. The variation for the two MX10-94M5 samples is also shown. Ideally, a clear genetic marker for passaging would emerge with the presence of a position in the genome where all of the passaged isolates shared the same genetic variant, and the variant would be distinct from the unpassaged variants; however this did not occur. There are six mutations associated with a change from unpassaged to passaged, but

they are only found in one isolate (00SMH279). For the MX10-94M5 sample, only two mutations among the dominant genotypes are observed (position 8755 and 8926).

With the availability of rare variants present in each passaged sample, it is possible to look at the genome positions where variation was observed in the dominant genotype and determine whether additional information is obtained by observing changes in the lower frequency mutations in the population. Mutations occurring at low frequency levels in the population are shown in Table 7 for the positions where a change was observed in the dominant genotype from 00SMH27 unpassaged to passaged. In several cases the unpassaged genotype was retained in the passaged sample as a rare variant indicating that even though the genomic locus is not explicitly identified as a functionally significant passage marker, determining the source of the passaged isolates could be improved by examining the identity of the rare variants.

Table 7. Sequence variation among dominant genotypes

Gene	Genome Position		MX10-94M5UP	00SMH279UP	00SMH279Vero	00SMH279HC6/36	MX10-94M5 Vero	Comments
Non-coding	28	Dominant	C	C	A	C	C	Evidence of high variability at this position
		Rare	-	-	-	T	T	
Non-coding	30	Dominant	Del	G	C	Del	Del	Single base deletion
NSP1	2454	Dominant	C	T	C	C	C	UP mutant retained as rare variant
		Rare	-	-		T	T	
NSP2	5493	Dominant	C	C	T	C	C	UP=C6/36
		Rare					T	
NSP2	5495	Dominant	C	C (Thr)	A (Asn)	C (Thr)	C	Non-synonymous change
NSP2	5502	Dominant	A	A	G	G	A	P strains differ from UP
		Rare					G	
NSP2	5520	Dominant	Del*	T	C	C	Del*	P strains differ from UP
NSP2	5523	Dominant	T	T	G	T	T	UP=C6/36
NSP4	7422	Dominant	C	C	T	T	C	P strains differ from UP
		Rare					T	

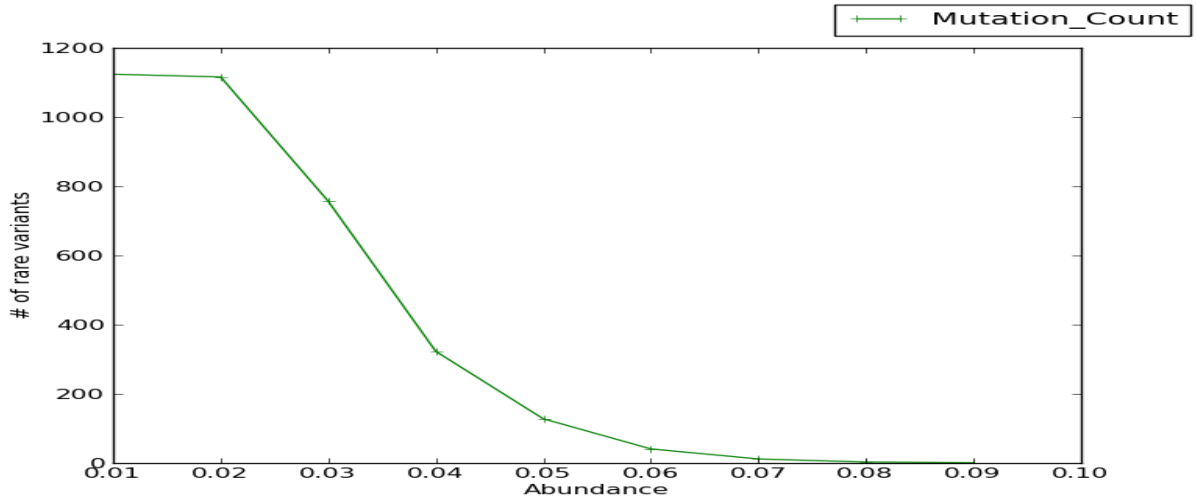
E1	10265	Dominant	T	C	T	T	T	P strains differ from UP
E1	10274	Dominant	G	T	C	C	G	UP mutant retained as rare variant
		Rare		-	T	T		

In searching for genetic markers of passage, one hypothesis is that genomic loci that differ in the natural isolates, but show a common genotype at low abundance levels in the passaged isolates, could be indicative of selection for a passage-associated genotype. To search for possible genetic markers of cell passage, we initially focused on mutations that differed in the dominant genotypes among the two natural isolates, which were highly divergent (75.8% identity over the genome). Table 8 shows that there were only 3 mutations that fit this criterion but in two of three cases, the mutation was synonymous making it less clear whether functional selection is in progress. Figure 7 shows the number of rare variants shared across all three passaged samples that occur at different abundance levels within each population. There are 1,120 rare variants, which were observed in all three passaged isolates, at 0.01% percent of the reads overlapping the genome position (abundance) or higher and just one mutation, which occurred in all three samples with at least 0.09% of the reads overlapping the genome position.

Table 8. Convergent rare variants. Shows the three mutations in the genome, where the two divergent isolates (00SMH279 and MX10-94M5) have different dominant genotypes but share the same rare variant in after passaged in cell culture. This table shows the associated gene, genome position, the dominant genotypes, the share rare genotype, and the abundance in the population of the rare variant for the three passaged samples (Vero 00SMH279, C6/36 00SMH279 and Vero MX10-94M5).

Gene	Genome Position	00SMH279	MX10-94M5	Rare Mutation	Vero 00SMH279 (frequency)	C6/36 00SMH279 (frequency)	Vero MX10-94M5 (frequency)
NSP1	576	GG**G** **G**	GG**A** G	GG**T** G	0.12	0.09	0.03
NSP2	3660	GA**C** D	GA**A** E	GA**T** D	0.04	0.03	0.08
E2	8810	CC**G** P	CC**T** P	CC**A** P	0.03	0.03	0.03

Figure 7. Abundance of rare variants found in all three passaged populations. X-axis shows abundance as the percentage of reads overlapping the position of interest and containing the rare variant. The y-axis counts the total number of rare variants found in all three samples with the minimum abundance on the x-axis.



A second set of candidate passage markers were selected using the frequency distribution in Figure 7 as a guide. Rare variants that occur in all three samples above a minimum abundance threshold were chosen. The threshold was set to 0.07% to target the 12 (collectively across all three samples) most abundant positions. The candidates are listed in Table 9 and indicate that 6 of the 12 mutations are non-synonymous changes. While many of the rare variants appear at very low frequency levels (< 0.1%), position 8896 stands out as a non-synonymous change in the E2 protein from Serine (S) to Leucine (L) that is relatively high in abundance (as high as 11.3% in the MX10 Vero cells) and could indicate that the genotype is selected in cell culture. Once we obtain the comparable unpassaged samples we can confirm the presence or absence of the mutations in the starting populations. If the absence of these 12 rare variant mutations in the unpassaged samples can be confirmed these positions should be considered carefully for marking evidence of cell culture passage.

Table 9. Abundant rare variants shared in cell culture. This table shows positions in the genome where all three passaged isolates have the same rare variants. Positions in red indicate non-synonymous mutations. The last three columns show the relative abundance as a percentage of overlapping mapped reads with the rare variant in each of the three passaged samples.

Gene	Genome Position	Dominant	Rare Mutation	Vero 00SMH279	C6/36 00SMH279	Vero MX10-94M5
NSP1	838	**C**AC **H**	**T**AC Y	0.1	0.08	0.07
NSP1	1131	CT**C** L	CT**T** L	0.08	0.09	0.08
NSP2	3018	AT**C** I	AT**T** I	0.08	0.07	0.08

NSP3	4666	**C**AC H	**T**AT Y	0.09	0.08	0.08
NSP3	5261	T**C**C S	T**T**C F	0.07	0.08	0.07
NSP4	6174	TG**C** C	TG**T** **C**	0.08	0.08	0.09
NSP4	7068	TC**C** C	TT**T** C	0.68	0.08	0.09
NSP4	7089	GT**G** V	GT**A** V	0.08	0.09	0.07
E2	8896	T**C**A S	T**T**A L	0.10	0.1	**11.3**
E2	9543	**C**CG P	**T**CG S	0.1	0.1	0.07
6K MP	9944	CC**T** P	CC**C** P	0.9	0.12	0.12
E1	10785	**G**AA E	**A**AA K	0.08	0.09	0.1

Analysis of human and equine monocyte replication patterns for VEEV strains

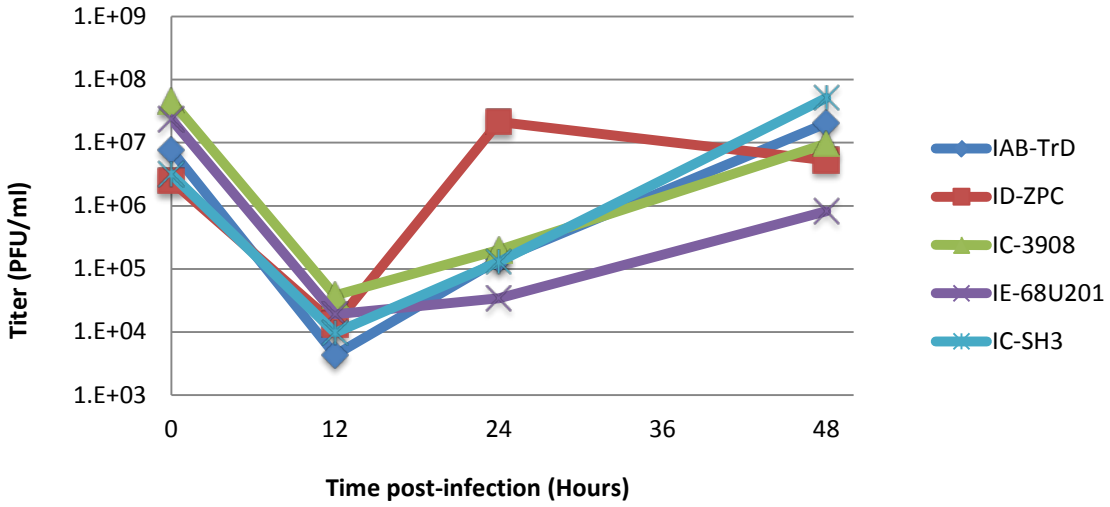
To determine human and equine monocyte replication patterns for genetically diverse VEEV strains, and subsequently correlate phenotypes with genetic signatures, the replication kinetics of 10 VEEV isolates representing major lineages and subtypes were to be examined in human and equine monocytes. Equine monocytes were to be collected from whole blood provided by donor horses at the Texas A&M College of Veterinary Medicine, and human monocytes provided by human donors at the UTMB blood bank.

Following unexpected delays in obtaining horse monocytes from our collaborators at Texas A&M University, it has become necessary to identify a new source for equine cellular material. We have not yet been able to finalize collaboration with a newly identified source, thus we are currently in pursuit of a new and more reliable source of horse blood for completion of this portion of the task during Year 3.

We performed the infection of human THP-1 macrophages using five VEEV strains: IAB TrD, IC 3908, IC SH3, ID ZPC738, and IE 68U201. Cells were infected at an MOI of 10, and harvested at 0, 12, 24 and 48 hours post-infection.

This preliminary experiment showed several interesting distinctions in replication patterns between the selected strains (Figure 8). Notably, the ID ZPC strain demonstrated increased replication between 12 and 24 hours, while the IE-68U201 strain replicated at a slower rate following 12 hours after infection. These results will be confirmed in subsequent experiments with additional harvesting time-points, and slightly modified infection conditions to eliminate the variability in titers at the initial infection time point (T0). Further, five additional VEEV strains will be included in the follow-up experiments.

Figure 8. VEEV growth kinetics in THP-1 macrophages



Discussion and Conclusions

Tools for rapid genotyping of equine encephalitis virus strains and elucidating their phylogenetic relationships are critically important for understanding why certain strains are likely to cause epizootic infection, and to enable us to forecast the incidence of potential epidemic events. The results above represent analysis of VEEV strains derived from a wide range of hosts and geographic regions. The collected data indicate that our microarray and sequencing-based genotyping tools effectively distinguish VEEV strains and allow us to cluster those strains according to their derivation and phenotypic history.

SNP-based phylogenetics

Our SNP analyses revealed that, in general, VEEV isolates do not group phylogenetically according to host, as strains clustered together were often derived from different host species. Nor could we identify a single mutation reliably associated with host; dozens of mutations were significantly associated with horse or human hosts, but to make host predictions of moderate accuracy required a combination of multiple mutations. This was not surprising for viruses capable of zoonotic infection that circulate among multiple hosts, as the host in which a strain was collected is somewhat a matter of chance.

We also observed instances in which individual isolates within a cluster had a distinct host deviation from other strains in the cluster. For instance, strains 63Z1 and MenaII were the only IE strains isolated from human infections. As these strains are nearly identical to isolates collected from hamsters and mosquitos, it is possible that these cases of human infection were due to variation in host phenotype or unusually high exposure to enzootic strains; the human host may have demonstrated an atypical propensity toward infection, resulting in disease being associated with a strain that is generally only observed in enzootic infection, rather than mutations in the VEEV strains.

Predicting genotype/phenotype associations was slightly more accurate based on variable positions from a whole genome alignment than based on kSNP SNPs, and models required fewer variations to achieve this accuracy. The VEEV genome is small enough that MSA of more than

100 sequences was possible. This is not usually the case for much larger bacterial genomes, making kSNP SNPs a good option for bacterial genotype/phenotype association studies.

Relying on non-random associations between subtype and sequence variation, we were able to build models to predict subtype. With 18-54 loci per subtype, prediction accuracy was 90-99%. However, accuracy was never 100% and strains that clustered phylogenetically with a different subtype were usually mis-labeled by the predictive models, indicating that we still do not have a good understanding of antigenic switch. For each of the types, there were alignment positions in which the allele was significantly associated with type, but at no positions was there perfect association, suggesting that serotype must result from more complex interactions at multiple positions. Previous investigators have explored mutations required for VEE to transition from the enzootic cycle (birds, small mammals, Culex mosquitos, forest habitats) to the epizootic cycle (Aedes/Ochlerotatus/Psorophora mosquitos, amplification in equids, transmission to humans). They (Anishchenko et al. 2006b) reported a single mutation in the E2 protein (T213 -> K or R) that, when engineered into a subtype ID enzootic strain, changed its serotype to IC and rendered it capable of causing viremia in horses, as well as infecting Ochlerotatus mosquitos. Our data suggest more complexity in this process, indicating that multiple loci are required to distinguish ID from IAB or IC serotypes.

Comparison of the phylogenetic tree predicted from whole genome SNPs was similar to that from whole genome multiple sequence alignment. Narrowing to single gene SNP trees showed that the E1 gene SNPs more closely represent the whole genome SNP tree than do the SNPs from the other envelope or capsid genes. This concurs with previous analyses based on sequence alignment rather than SNPs (Bendy et al. 1964). However, these results emphasize that use of a small region of the genome for SNP analysis provides lower resolution than whole genome SNPs, and with some genes even results in different tree topology. A whole genome SNP approach more effectively represents complete phylogenetic relationships to reveal distinctions that otherwise would be overlooked.

There are limitations of a k-mer based approach to SNP discovery compared with full sequence alignment, particularly for highly variable RNA viruses. SNPs in close proximity (within the k-mer context around the SNP) or identical k-mer context in non-homologous regions among genomes cause errors in loci identification. Nonetheless, our comparison of data derived from multiple sequence alignments versus SNP analysis revealed that the resultant trees were very similar and reliably identified comparable splits. These observed similarities are important in that they support the use of our unique SNP genotyping tools as a cost-efficient, fast method for characterizing samples without available sequence data using SNP arrays.

Rapid microarray-based strain genotyping

We have shown here that data obtained from SNP arrays are capable of reliably clustering strains in accordance with their respective whole genome sequence data. This technology would be particularly useful for rapidly evaluating a novel strain from an epizootic outbreak event. We demonstrated that we could accurately SNP type an isolate directly from RNA extracted from infected host tissue without isolation or culturing, a huge advantage in surveillance efforts where field and medical personnel need rapid strain characterization and may not have access to containment labs for culturing or the resources and time required for genome sequencing. When comparing SNP allele calls from VEEV SNP microarray versus whole or draft sequences, we found an overall concordance rate of 97.5%. Our previous studies have

shown that the SNP array calls had a higher concordance rate with finished genomes (99.8%) versus draft genomes (95.5%) (Gardner et al. 2013). Due to high mutation rates, loci are missing in many VEEV strains due to multiple nucleotide differences and indels, making this virus particularly challenging for SNP array classification. In the future, we hope to improve our methods to detect missing loci. Despite the noted limitations, placement on the phylogeny was very accurate for clustering array and genome data by subtype, host, country, and year.

Sequence markers of laboratory passage

Finally, we performed deep sequencing of strains that had undergone passage through cell culture and compared these data to unpassaged sequence data, with the goal of identifying a dominant polymorphic genetic marker indicative of passaging. As is described above, a clear dominant marker was not observed in our analysis. There were, however, 1000+ rare variants observed in passaged isolates above our set thresholds. These rare variants were explored further to identify whether they might serve as useful markers of passage. The pool of markers was downselected to six non-synonymous changes, with one in particular being observed at higher abundance. These markers may serve as potential indicators of genotypes associated with culture, a possibility which will be confirmed in further analysis once the complete set of passaged and unpassaged strains are available.

We have demonstrated in our studies thus far that the use of our novel SNP analysis tools and microarrays can effectively characterize VEEV strains in a rapid and highly accurate manner. In addition, once genetic markers for culture passage are further explored, these tools may be capable of distinguishing between those strains which have undergone laboratory amplification. In combination, these data support our approach as a promising tool for understanding the progression and incidence of VEEV infection.

References

Aguilar PV, Greene IP, Coffey LL, Medina G, Moncayo AC, Anishchenko M, Ludwig GV, Turell MJ, O'Guinn ML, Lee J et al. 2004. Endemic Venezuelan equine encephalitis in northern Peru. *Emerging infectious diseases* **10**(5): 880-888.

Anishchenko M, Bowen RA, Paessler S, Austgen L, Greene IP, Weaver SC. 2006a. Venezuelan encephalitis emergence mediated by a phylogenetically predicted viral mutation. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 4994-4999.

-. 2006b. Venezuelan encephalitis emergence mediated by a phylogenetically predicted viral mutation. *Proceedings of the National Academy of Sciences of the United States of America* **103**(13): 4994-4999.

Bendy RH, Jr., Nuccio PA, Wolfe E, Collins B, Tamburro C, Glass W, Martin CM. 1964. Relationship of Quantitative Wound Bacterial Counts to Healing of Decubiti: Effect of Topical Gentamicin. *Antimicrob Agents Chemother (Bethesda)* **10**: 147-155.

Bronze MS, Huycke MM, Machado LJ, Voskuhl GW, Greenfield RA. 2002. Viral agents as biological weapons and agents of bioterrorism. *Am J Med Sci* **323**(6): 316-325.

Chen-Harris H, Borucki MK, Torres C, Slezak TR, Allen JE. 2013. Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs. *BMC Genomics* **14**: 96.

Ferro C, Olano VA, Ahumada M, Weaver S. 2008. [Mosquitos (Diptera: Culicidae) in the small village where a human case of Venezuelan equine encephalitis was recorded]. *Biomedica* **28**(2): 234-244.

Forshey BM, Guevara C, Laguna-Torres VA, Cespedes M, Vargas J, Gianella A, Vallejo E, Madrid C, Aguayo N, Gotuzzo E et al. Arboviral etiologies of acute febrile illnesses in Western South America, 2000-2007. *PLoS Negl Trop Dis* **4**(8): e787.

Gardner S, Slezak T. 2010. Scalable SNP Analyses of 100+ Bacterial or Viral Genomes. *J Forensic Res* **1**: 107, doi:110.4172/2157-7145.1000107.

Gardner SN, Thissen J, McLoughlin K, Slezak T, Jaing C. 2013. Optimizing SNP microarray probe design for high accuracy microbial genotyping. *J Microbio Meth*: http://dx.doi.org/10.1016/j.mimet.2013.1007.1006.

Hall BG. 2014. SNP-Associations and Phenotype Predictions from Hundreds of Microbial Genomes without Genome Alignments. *PloS one* **9**(2): e90490.

Hawley RJ, Eitzen EM, Jr. 2001. Biological weapons--a primer for microbiologists. *Annu Rev Microbiol* **55**: 235-253.

Jaing C, Gardner S, McLoughlin K, Mulakken N, Alegria-Hartman M, Banda P, Williams P, Gu P, Wagner M, Manohar C et al. 2008. A functional gene array for detection of bacterial virulence elements. *PLoS One* **3**(5): e2163.

Johnson KM, Shelokov A, Peralta PH, Dammin GJ, Young NA. 1968. Recovery of Venezuelan equine encephalomyelitis virus in Panama. A fatal case in man. *The American journal of tropical medicine and hygiene* **17**(3): 432-440.

Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**: 980-984.

Quiroz E, Aguilar PV, Cisneros J, Tesh RB, Weaver SC. 2009. Venezuelan equine encephalitis in Panama: fatal endemic disease and genetic diversity of etiologic viral strains. *PLoS Negl Trop Dis* **3**(6): e472.

Venkatachalam B, Apple J, St John K, Gusfield D. 2010. Untangling tanglegrams: comparing trees by their drawings. *IEEE/ACM Trans Comput Biol Bioinform* **7**: 588-597.

Vilcarromero S, Aguilar PV, Halsey ES, Laguna-Torres VA, Razuri H, Perez J, Valderrama Y, Gotuzzo E, Suarez L, Cespedes M. 2010. Venezuelan equine encephalitis and 2 human deaths. *Peru Emerg Infect Dis* **16**: 553-556.

Vilcarromero S, Aguilar PV, Halsey ES, Laguna-Torres VA, Razuri H, Perez J, Valderrama Y, Gotuzzo E, Suarez L, Cespedes M et al. Venezuelan equine encephalitis and 2 human deaths, Peru. *Emerging infectious diseases* **16**(3): 553-556.

Vilcarromero S, Laguna-Torres VA, Fernandez C, Gotuzzo E, Suarez L, Cespedes M, Aguilar PV, Kochel TJ. 2009. Venezuelan equine encephalitis and upper gastrointestinal bleeding in child. *Emerging infectious diseases* **15**(2): 323-325.

Watts DM, Callahan J, Rossi C, Oberste MS, Roehrig JT, Wooster MT, Smith JF, Cropp CB, Gentrau EM, Karabatsos N et al. 1998. Venezuelan equine encephalitis febrile cases among humans in the Peruvian Amazon River region. *The American journal of tropical medicine and hygiene* **58**(1): 35-40.

Watts DM, Lavera V, Callahan J, Rossi C, Oberste MS, Roehrig JT, Cropp CB, Karabatsos N, Smith JF, Gubler DJ et al. 1997. Venezuelan equine encephalitis and Oropouche virus

infections among Peruvian army troops in the Amazon region of Peru. *Am J Trop Med Hyg* **56**(6): 661-667.
Weaver SC, Barrett AD. 2004. Transmission cycles, host range, evolution and emergence of arboviral disease. *Nature reviews Microbiology* **2**(10): 789-801.
Weaver SC, Ferro C, Barrera R, Boshell J, Navarro JC. 2004. Venezuelan equine encephalitis. *Annu Rev Entomol* **49**: 141-174.
Weaver SC, Reisen WK. 2009. Present and future arboviral threats. *Antiviral Res* **85**(2): 328-345.